IDENTIFYING PHYLOGENETIC TREES

M Bordewich, K. T. Huber and C Semple

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Report Number: UCDMS2004/7 MAY 2004

IDENTIFYING PHYLOGENETIC TREES

MAGNUS BORDEWICH, KATHARINA T. HUBER, AND CHARLES SEMPLE

ABSTRACT. A central problem that arises in evolutionary biology is that of displaying partitions of subsets of a finite set X on a tree whose vertices are partially labelled with the elements of X. Such a tree is called an X-tree and, for a collection $\mathcal{C}$ of partitions of subsets of X, characterisations for the existence and uniqueness of an X-tree that displays $\mathcal{C}$ have been previously given in terms of chordal graphs. In this paper, we obtain two closely related characterisations also in terms of chordal graphs. The first describes when $\mathcal{C}$ identifies an X-tree, and the second describes when a compatible subset of $\mathcal{C}$ is of maximum size.

1. INTRODUCTION

For a finite set X, an *X-tree* $\mathcal{T} = (T; \phi)$ is an ordered pair consisting of a tree T, with vertex set V say, and a map $\phi : X \to V$ with the property that, for all $v \in V$ with degree at most two, $v \in \phi(X)$. An X-tree is also called a *semi-labelled tree*. If ϕ is a bijection from X into the leaf set of T, then $\mathcal{T}$ is a *phylogenetic X-tree*. In Fig. 1 (ignoring the two bold edges), $\mathcal{T}_1$ and $\mathcal{T}_2$ are both X-trees, where $\mathcal{T}_2$ is also phylogenetic. The set X is called the *label set* of $\mathcal{T}$ and is denoted $\mathcal{L}(\mathcal{T})$. Furthermore, if v is a vertex of T, then $\phi^{-1}(v)$ is the *label set* of v, and the elements of this set are the elements of X *labelling* v. A *character on X* is a function χ from a non-empty subset X' of X into a set C of *character states*. If $|C| = 2$, then χ is a *two-state* character. For our purposes, the elements of C are not important. The real importance is the partition of X' induced by χ. To this end, we let $\pi(\chi)$ denote the partition of X' corresponding to $\{\chi^{-1}(\alpha) : \alpha \in C\}$.

Let χ be a character on X and let $\mathcal{T} = (T; \phi)$ be an X-tree. We say that $\mathcal{T}$ *displays* χ if there is a subset E of edges of T such that, for all $A, B \in \pi(\chi)$, $\phi(A)$ and $\phi(B)$ are subsets of the vertex sets of different components of the graph obtained from T by deleting the edges in E. Extending the examples of X-trees shown in Fig. 1, let $\chi : \{a, c, f, i\} \to \{0, 1\}$ be the character on X defined by setting $\chi(x) = 0$ for each $x \in \{a, c\}$, and $\chi(x) = 1$ for each $x \in \{f, i\}$. Then $\mathcal{T}_1$ displays

Date: 30 April 2004.

1991 *Mathematics Subject Classification.* 05C05; 92D15.

Key words and phrases. Phylogenetic tree, Identifying, Restricted chordal completion.

The first author was supported by the New Zealand Institute of Mathematics and its Applications funded programme *Phylogenetic Genomics*. The second author was supported by The Swedish Research Council (VR) and the third author was supported by the New Zealand Marsden Fund (UOC310). All authors thank The Swedish Foundation for International Cooperation in Research and Education (STINT).

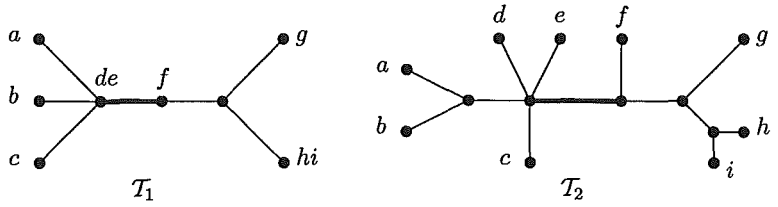

FIGURE 1. An X-tree $\mathcal{T}_1$, and a phylogenetic X-tree $\mathcal{T}_2$, where $X = \{a, b, c, d, e, f, g, h, i\}$.

χ. More generally, $\mathcal{T}$ *displays* a collection $\mathcal{C}$ of characters on X if $\mathcal{T}$ displays each character in $\mathcal{C}$, in which case $\mathcal{C}$ is *compatible.*

In evolutionary biology, phylogenetic trees are used to represent the evolutionary relationships of a collection of present-day species. X-trees are a convenient mathematical generalisation of phylogenetic trees. The data that is typically used to reconstruct phylogenetic trees is based on characters, where a character assigns two species the same character state if they share the corresponding feature. Given a collection $\mathcal{C}$ of characters, a central problem in phylogenetics is to determine if there is a semi-labelled tree that displays this collection. Commonly known as the *character compatibility problem* or the *perfect phylogeny problem*, this problem in general is NP-complete [1, 8]. However, there are attractive characterisations for this existence problem and its corresponding uniqueness problem in terms of chordal graphs.

In this paper, we present two related characterisations, again in terms of chordal graphs. In practice, even if there is a semi-labelled tree that displays $\mathcal{C}$, it is unlikely that $\mathcal{C}$ *defines* it; that is, no other semi-labelled tree displays $\mathcal{C}$. However, a closely related notion, and one that is almost as good, is that of "identifiability". Associated with each edge e of an X-tree $\mathcal{T} = (T; \phi)$ is an X-*split*; that is, a bipartition of X into the label sets of the two connected components of $\mathcal{T}\backslash e = (T\backslash e, \phi)$. An X-tree $\mathcal{T}'$ is a *refinement* of $\mathcal{T}$ if every X-split of $\mathcal{T}$ is an X-split of $\mathcal{T}'$. Intuitively, $\mathcal{T}'$ is a refinement of $\mathcal{T}$ if $\mathcal{T}$ can be obtained from $\mathcal{T}'$ by contracting edges. In Fig. 1, $\mathcal{T}_2$ is a refinement of $\mathcal{T}_1$. We say that $\mathcal{C}$ *identifies* an X-tree $\mathcal{T}$ if $\mathcal{T}$ displays $\mathcal{C}$ and every X-tree $\mathcal{T}'$ that displays $\mathcal{C}$ is a refinement of $\mathcal{T}$. Our first characterisation (Theorem 1.3) describes when a collection of characters identifies an X-tree in terms of chordal graphs.

As one might expect, biological data can often be inconsistent and so, for a collection $\mathcal{C}$ of characters there may be no phylogenetic tree that displays $\mathcal{C}$. Thus a natural problem is to determine a maximum-sized subset $\mathcal{C}'$ of $\mathcal{C}$ for which there is a phylogenetic tree that displays $\mathcal{C}'$. Of course, since the existence problem is NP-complete, this problem is NP-hard. But, like the characterisations mentioned above, there is a characterisation of this problem in terms of chordal graphs. This is our second characterisation (Theorem 1.5).

The rest of this section contains some necessary preliminaries, background material, and the statements of Theorems 1.3 and 1.5. The next section shows that

the conditions in Theorem 1.3 cannot be weakened. Section 3 contains the proof of the sufficiency part of Theorem 1.3 in the restricted setting of two-state characters, which is then used to prove Theorem 1.3 for arbitrary characters in Section 4. Section 5 contains the proof of Theorem 1.5. Throughout the paper, the notation and terminology follows Semple and Steel [7] with one exception. This exception is that we say "$\mathcal{T}$ displays $\mathcal{C}$" instead of "$\mathcal{C}$ is convex on $\mathcal{T}$". Furthermore, for an X-tree $\mathcal{T} = (T; \phi)$, we will often refer to the vertices and edges of T as the vertices and edges of $\mathcal{T}$ provided no ambiguity arises. Let $\psi : A \to B$ be a map and let $b \in B$, we will frequently use $\psi^{-1}(b)$ to denote the (possibly empty) subset of A whose elements are mapped to b under ψ. Lastly, for an interesting and easy reading discussion of the perfect phylogeny problem, we refer the reader to [9].

Let $\mathcal{C}$ be a collection of characters on X and let $\mathcal{T} = (T; \phi)$ be an X-tree. Let $X', X'' \subseteq X$. The minimal subtree of T that connects the vertices of $\phi(X')$ is denoted by $\mathcal{T}(X')$. Furthermore, we say that $\mathcal{T}(X') \cap \mathcal{T}(X'')$ is *non-empty* if the vertex sets of $\mathcal{T}(X')$ and $\mathcal{T}(X'')$ are disjoint. We now define two graphs each of which has vertex set

$$V(\mathcal{C}) = \bigcup_{\chi \in \mathcal{C}} \{(\chi, A) : A \in \pi(\chi)\}.$$

(i) The *partition intersection graph of* $\mathcal{C}$, denoted $\mathrm{int}(\mathcal{C})$, is the graph that has vertex set $V(\mathcal{C})$ and an edge joining (χ_1, A) and (χ_2, B) if $A \cap B$ is non-empty.
(ii) The *subtree intersection graph of* $\mathcal{T}$ *induced by* $\mathcal{C}$, denoted $\mathrm{int}(\mathcal{C}, \mathcal{T})$, is the graph that has vertex set $V(\mathcal{C})$ and an edge joining (χ_1, A) and (χ_2, B) if $\mathcal{T}(A) \cap \mathcal{T}(B)$ is non-empty.

A graph is *chordal* if every cycle that contains at least four vertices has an edge connecting two non-consecutive vertices. A graph G is a *restricted chordal completion* of $\mathrm{int}(\mathcal{C})$ if G is a chordal graph that can be obtained from $\mathrm{int}(\mathcal{C})$ by adding only edges that join vertices whose first components are different. It is well-known that if $\mathcal{T}$ displays $\mathcal{C}$, then $\mathrm{int}(\mathcal{C}, \mathcal{T})$ is a restricted chordal completion of $\mathrm{int}(\mathcal{C})$ (for example, see [6]). For example, let $\mathcal{C}$ be the set of characters $\{ab|ce, cd|af, bd|ef\}$, where $ab|ce$ denotes a character χ such that $\pi(\chi) = \{\{a, b\}, \{c, e\}\}$. Then the partition intersection graph of $\mathcal{C}$ and an associated restricted chordal completion are given in Fig. 2. Also, the X-tree $\mathcal{T}$ shown in Fig. 2 displays $\mathcal{C}$, and $\mathrm{int}(\mathcal{C}, \mathcal{T})$ is the restricted chordal completion of $\mathrm{int}(\mathcal{C})$ shown.

Theorem 1.1 is a graph-theoretic characterisation for when there exists an X-tree that displays a given collection of characters. This result is indicated in [2] and [5], and formally proved in [8].

Theorem 1.1. *Let $\mathcal{C}$ be a collection of characters on X. Then there exists an X-tree that displays $\mathcal{C}$ if and only if there exists a restricted chordal completion of* $\mathrm{int}(\mathcal{C})$.

To describe the uniqueness analogue of Theorem 1.1, we require some further definitions. Let χ be a character on X and let $\mathcal{T}$ be an X-tree. Let e be an edge of $\mathcal{T}$. We say that χ *distinguishes* e if every set of edges of $\mathcal{T}$ that displays χ contains

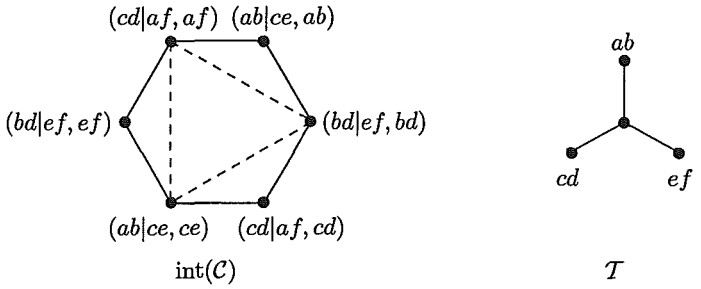

FIGURE 2. The partition intersection graph of the set of characters $\mathcal{C} = \{ab|ce, cd|af, bd|ef\}$ (solid lines), a restricted chordal completion of $\text{int}(\mathcal{C})$ (dashed lines), and an X-tree that displays $\mathcal{C}$ where $\text{int}(\mathcal{C}, \mathcal{T})$ is given by this restricted chordal completion.

e. Intuitively, this means that if e is contracted in $\mathcal{T}$, then the resulting X-tree does not display χ. For example, in Fig. 1, the bold edge in $\mathcal{T}_2$ is distinguished by the character $\chi : \{a, c, f, i\} \to \{0, 1\}$, where $\chi(x) = 0$ for each $x \in \{a, c\}$ and $\chi(x) = 1$ for each $x \in \{f, i\}$. We say that $\mathcal{T}$ is *distinguished* by a collection $\mathcal{C}$ of characters on X if each edge of $\mathcal{T}$ is distinguished by an element in $\mathcal{C}$. A restricted chordal completion G of $\text{int}(\mathcal{C})$ is *minimal* if, for every non-empty subset F of $E(G) - E(\text{int}(\mathcal{C}))$, the graph $G \backslash F$ is not chordal. Lastly, a phylogenetic X-tree is *binary* if every interior vertex has degree three. The following theorem is established in [6].

Theorem 1.2. *Let $\mathcal{C}$ be a collection of characters on X. Then there is an unique X-tree that displays $\mathcal{C}$ if and only if the following conditions hold:*

(i) *there is a binary phylogenetic X-tree that displays $\mathcal{C}$ and is distinguished by $\mathcal{C}$; and*
(ii) *there is a unique minimal restricted chordal completion of* $\text{int}(\mathcal{C})$.

Moreover, if $\mathcal{T}$ is the unique X-tree displaying $\mathcal{C}$, then $\mathcal{T}$ satisfies the properties in (i) and $\text{int}(\mathcal{C}, \mathcal{T})$ *is the unique minimal restricted chordal completion of* $\text{int}(\mathcal{C})$.

To describe Theorem 1.3, a characterisation of identifiability in terms of chordal graphs, we first need a stronger notion of distinguish. Let $\mathcal{T} = (T; \phi)$ be an X-tree and let $e = \{u, v\}$ be an edge of T. Then e is *strongly distinguished* by a character χ on X, if there exist A and B in $\pi(\chi)$ such that the following hold:

(i) $\phi(A)$ is a subset of the vertex set of the component of $T \backslash e$ containing u;
(ii) the vertex set of each component of $T \backslash u$, except for the one containing v, contains an element of $\phi(A)$;
(iii) $\phi^{-1}(u)$ is a subset of A;
(iv) $\phi(B)$ is a subset of the vertex set of the component of $T \backslash e$ containing v;
(v) the vertex set of each component of $T \backslash v$, except for the one containing u, contains an element of $\phi(B)$;

(vi) $\phi^{-1}(v)$ is a subset of B.

This definition of strongly distinguished extends the definition given for phylogenetic trees in [6]. To illustrate strongly distinguished, in Fig. 1, the bold edge in $\mathcal{T}_1$ is strongly distinguished by the character $\chi : \{a, b, c, d, e, f, i\} \to \{0, 1\}$, where $\chi(x) = 0$ for each $x \in \{a, b, c, d, e\}$ and $\chi(x) = 1$ for each $x \in \{f, i\}$. We say $\mathcal{T}$ is *strongly distinguished* by a collection $\mathcal{C}$ of characters if every edge of $\mathcal{T}$ is strongly distinguished by some character in $\mathcal{C}$. Observe that if e is strongly distinguished by a character χ, then it is also distinguished by χ. However, the converse does not hold. Furthermore, the X-split induced by an edge e of $\mathcal{T}$ strongly distinguishes e.

For a collection $\mathcal{C}$ of characters on X, let $\mathcal{G}(\mathcal{C})$ denote the set of graphs

$$\mathcal{G}(\mathcal{C}) = \{G : \text{there is an } X\text{-tree } \mathcal{T} \text{ displaying } \mathcal{C} \text{ with } G = \text{int}(\mathcal{C}, \mathcal{T})\}.$$

Observe that $\mathcal{G}(\mathcal{C})$ is a subset of the collection of all restricted chordal completions of $\mathcal{C}$. A useful partial order $\leq$ on $\mathcal{G}(\mathcal{C})$ is obtained by setting, for all $G_1, G_2 \in \mathcal{G}(\mathcal{C})$, $G_1 \leq G_2$ if the edge set of G_1 is a subset of the edge set of G_2. Furthermore, for a compatible collection $\mathcal{C}$ of characters on X, we say that $\mathcal{C}$ *infers* a character χ if every X-tree that displays $\mathcal{C}$ also displays χ.

Theorem 1.3. *Let $\mathcal{C}$ be a collection of characters on X. Then $\mathcal{C}$ identifies an X-tree if and only if the following conditions hold:*

(i) *there is an X-tree that displays $\mathcal{C}$ and, for every edge e of this tree, there is a character on X inferred by $\mathcal{C}$ that strongly distinguishes e; and*
(ii) *there is a unique maximal element in $\mathcal{G}(\mathcal{C})$.*

Moreover, if $\mathcal{C}$ identifies an X-tree $\mathcal{T}$, then $\mathcal{T}$ satisfies the properties in (i) and $\text{int}(\mathcal{C}, \mathcal{T})$ *is the unique maximal element of $\mathcal{G}(\mathcal{C})$.*

Observe that if $\mathcal{C}$ identifies a binary phylogenetic X-tree $\mathcal{T}$, then $\mathcal{T}$ is the unique X-tree that displays $\mathcal{C}$. Also note that, despite Theorem 1.2, examples can be found such that $\mathcal{C}$ identifies an X-tree, but there is not a unique minimal restricted chordal completion.

Before describing Theorem 1.5, we present an attractive consequence of Theorem 1.3 for collections of semi-labelled trees.

Let X' be a subset of X. An X-tree $\mathcal{T}$ *displays* an X'-tree $\mathcal{T}'$ if the X'-tree obtained from $\mathcal{T}(X')$ by suppressing any unlabelled vertex of degree two is a refinement of $\mathcal{T}'$. In general, $\mathcal{T}$ *displays* a collection $\mathcal{P}$ of semi-labelled trees if $\mathcal{T}$ displays every tree in $\mathcal{P}$. Let $\mathcal{L}(\mathcal{P})$ denote the union of the label sets of the trees in $\mathcal{P}$. Analogous to the corresponding definition for a collection of characters, a collection $\mathcal{P}$ of semi-labelled trees *identifies* a semi-labelled tree $\mathcal{T}$ with label set $\mathcal{L}(\mathcal{P})$ if $\mathcal{T}$ displays $\mathcal{P}$ and all semi-labelled trees with label set $\mathcal{L}(\mathcal{P})$ that display $\mathcal{P}$ are refinements of $\mathcal{T}$.

Let $\mathcal{T} = (T; \phi)$ be an X-tree and let e be an edge of T. Suppose that $A|B$ is an X-split of $\mathcal{T}$ induced by an edge e. We define the two-state character χ_e on X by setting $\chi_e(x) = 0$ if $x \in A$ and $\chi_e(x) = 1$ if $x \in B$. Let $\mathcal{C}(\mathcal{T})$ denote the collection of

all two-state characters on X that are obtained in this way by the edges of $\mathcal{T}$. The following corollary is an easy consequence of [8, Proposition 2(2)] and Theorem 1.3.

Corollary 1.4. *Let $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be a collection of semi-labelled trees and let $X = \mathcal{L}(\mathcal{P})$. Let $\mathcal{C} = \bigcup_{i=1}^{n} \mathcal{C}(\mathcal{T}_i)$, then $\mathcal{P}$ identifies an X-tree $\mathcal{T}$ if and only if:*

(i) *there is an X-tree that displays $\mathcal{P}$ and, for every edge e of this tree, there is a character on X inferred by $\mathcal{C}$ that strongly distinguishes e; and*
(ii) *there is a unique maximal element in $\mathcal{G}(\mathcal{C})$.*

Moreover, if $\mathcal{P}$ identifies an X-tree $\mathcal{T}$, then $\mathcal{T}$ satisfies the properties in (i) and $\mathrm{int}(\mathcal{C}, \mathcal{T})$ *is the unique maximal element of $\mathcal{G}(\mathcal{C})$.*

To the best of our knowledge, the characterisations described in Theorem 1.3 and Corollary 1.4 are the first non-trivial characterisations for when a collection of characters identifies an X-tree and for when a collection of semi-labelled trees identifies an X-tree. In regards to the latter characterisation, we note here that Daniel [3] has described a polynomial-time algorithm for recognising if a given collection of "rooted" phylogenetic trees identifies a rooted phylogenetic tree.

Lastly, we state Theorem 1.5. Let $\mathcal{C}$ be a collection of characters on X. A graph G is a *chordal completion* of $\mathrm{int}(\mathcal{C})$ if G is a chordal graph that can be obtained from $\mathrm{int}(\mathcal{C})$ by adding edges. Unlike a restricted chordal completion of $\mathrm{int}(\mathcal{C})$, there is no restriction on how these edges are added. Furthermore, $\mathrm{int}(\mathcal{C})$ always has a chordal completion; simply add edges between every pair of non-adjacent vertices. Now let G be a chordal completion of $\mathrm{int}(\mathcal{C})$. A character χ in $\mathcal{C}$ is *broken in G* if there exist distinct $A, A' \in \pi(\chi)$ such that (χ, A) and (χ, A') are joined by an edge in G. A subset $\mathcal{C}'$ of $\mathcal{C}$ is *broken* if there exists a chordal completion of $\mathrm{int}(\mathcal{C})$ in which the broken characters are precisely the elements of $\mathcal{C}'$.

Theorem 1.5. *Let $\mathcal{C}$ be a collection of characters on X and let $\mathcal{C}'$ be a subset of $\mathcal{C}$. Then $\mathcal{C}'$ is a maximum-sized compatible subset of $\mathcal{C}$ if and only if $\mathcal{C} - \mathcal{C}'$ is a minimum-sized broken subset of $\mathcal{C}$.*

2. Some Enlightening Examples

The main purpose of this section is to show that the necessary and sufficient conditions in the statement of Theorem 1.3 cannot be weakened. Simply for reasons of convenience, in this section we view a character on X as a partition of a subset of X. Also, we have labelled the vertex (χ, A) in $\mathrm{int}(\mathcal{C})$ by A only, if there is no other character $\chi' \in \mathcal{C}$ with $A \in \pi(\chi')$.

Firstly, it is not sufficient for only (i) to hold in the statement of Theorem 1.3. To see this, let $\mathcal{C} = \{ab|ce, cd|af, bd|ef\}$ be a collection of characters. Then the semi-labelled tree $\mathcal{T}_1$ shown in Fig. 3 displays $\mathcal{C}$ and every edge of $\mathcal{T}_1$ is strongly distinguished by $\mathcal{C}$. However, as $\mathcal{T}_2$ also displays $\mathcal{C}$, it is easily checked that $\mathcal{C}$ does not identify any X-tree.

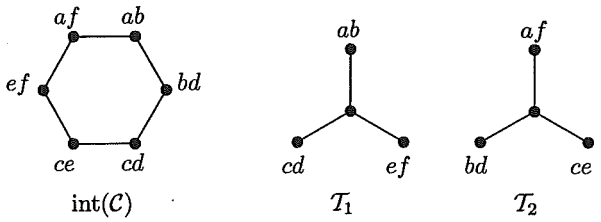

FIGURE 3. Two trees displaying $\mathcal{C} = \{ab|ce, cd|af, bd|ef\}$, both strongly distinguished by $\mathcal{C}$ and neither a refinement of the other.

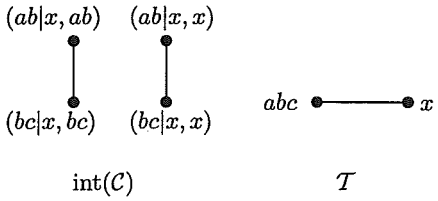

FIGURE 4. The intersection graph of, and the tree identified by $\mathcal{C} = \{ab|x, bc|x\}$.

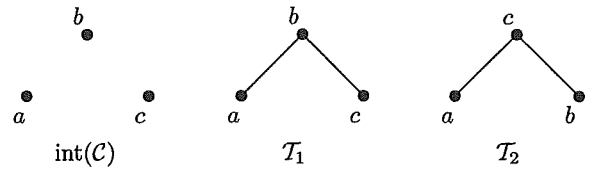

FIGURE 5. The intersection graph of, and two trees displaying $\mathcal{C} = \{a|b|c\}$.

The second example shows that in the statement of Theorem 1.3 we need only a unique maximal element of $\mathcal{G}(\mathcal{C})$: requiring a unique maximal restricted chordal completion of $\text{int}(\mathcal{C})$ would be too strong. This contrasts with the second condition in Theorem 1.2, however a *minimal* restricted chordal completion of $\text{int}(\mathcal{C})$ is always element of $\mathcal{G}(\mathcal{C})$. Let $\mathcal{C} = \{ab|x, bc|x\}$ be a collection of characters on $\{a, b, c, x\}$; it is easily checked that $\mathcal{C}$ identifies the semi-labelled tree $\mathcal{T}$, shown in Fig. 4. However, the partition intersection graph $\text{int}(\mathcal{C})$, also shown in Fig. 4, has two maximal restricted chordal completions. A routine check shows that adding an edge between $(ab|x, ab)$ and $(bc|x, x)$ creates one maximal restricted chordal completion of $\text{int}(\mathcal{C})$; another can be obtained by adding an edge between $(bc|x, bc)$ and $(ab|x, x)$.

Finally, by considering the collection $\mathcal{C} = \{a|b|c\}$, it is easily seen that we cannot weaken "strongly distinguish" in Theorem 1.3 to "distinguish". Fig. 5 shows $\text{int}(\mathcal{C})$, which is its own unique restricted chordal completion, and two trees which display $\mathcal{C}$ in which every edge is distinguished.

3. Sufficiency of Theorem 1.3 for Two-State Characters

In this section, we prove the sufficiency part of Theorem 1.3 for a collection of two-state characters. In particular, we prove the following theorem.

Theorem 3.1. *Let $\mathcal{C}$ be a collection of two-state characters on X. Then $\mathcal{C}$ identifies an X-tree if the following conditions hold:*

(i) *there is an X-tree that displays $\mathcal{C}$ and, for every edge e of this tree, there is a character on X inferred by $\mathcal{C}$ that strongly distinguishes e; and*
(ii) *there is a unique maximal element in $\mathcal{G}(\mathcal{C})$.*

To establish Theorem 3.1, we first prove two lemmas.

Lemma 3.2. *Let $\mathcal{C}$ be a collection of two-state characters on X, and let $\mathcal{T}$ be an X-tree that displays $\mathcal{C}$ and is strongly distinguished by $\mathcal{C}$. Let $\mathcal{T}'$ be an X-tree that displays $\mathcal{C}$. If $\mathrm{int}(\mathcal{C},\mathcal{T}') \leq \mathrm{int}(\mathcal{C},\mathcal{T})$, then $\mathcal{T}'$ is a refinement of $\mathcal{T}$.*

Proof. Let $\mathcal{T} = (T;\phi)$ and $\mathcal{T}' = (T';\phi')$. The proof is by induction on the size of X. Clearly, the result holds if $|X| \in \{1,2\}$. Now let $|X| = n$ and assume that the lemma holds for when $|X|$ is at most $n-1$, where $n \geq 3$. We consider two cases depending upon whether

(i) $\mathcal{T}$ contains a multiply-labelled leaf v, or
(ii) all leaves of $\mathcal{T}$ are singularly-labelled, in which case there is a vertex u of $\mathcal{T}$ which has at most one adjacent vertex that is not a leaf.

Depending on which case occurs, let L denote the following label sets: in case (i), L denotes the label set of v; and in case (ii), L denotes the union of the label sets of u and each of the leaves adjacent to u. Observe that $L|(X-L)$ is an X-split of $\mathcal{T}$ where $X-L$ may be possibly empty. If $X-L$ is non-empty, let e denote the edge of $\mathcal{T}$ that induces this X-split. We next show that, provided $X-L$ is non-empty, there must be an edge e' in $\mathcal{T}'$ such that the X-split induced by e' is $L|(X-L)$.

Let $\{x,y\}$ be the edge in T' that induces the X-split $A|X-A$ such that $L \subseteq A$ and $|A|$ is minimised. Suppose that $A-L$ is non-empty, and consider T. By starting at one end-vertex of e in T, and continually moving along the edges of T towards a subtree of T that contains labels from both $A-L$ and $X-A$, we eventually find an edge $\{z,w\}$ of T on this path such that each component of $T\backslash w$ which does not contain z, contains only labels from $A-L$ or only labels from $X-A$, and no previously traversed edge on this path has this property. Since the edge $\{z,w\}$ is strongly distinguished by $\mathcal{C}$, there is a character χ in $\mathcal{C}$ that strongly distinguishes $\{z,w\}$. This implies that there is an element $A' \in \pi(\chi)$ with $A' \cap L = \emptyset$ such that both $A' \cap (A-L)$ and $A' \cap (X-A)$ are non-empty. It now follows that $\mathcal{T}'(A')$ contains both x and y, and, since $|A|$ is minimised, $\mathcal{T}'(L)$ must contain either x or y. Hence, $\mathcal{T}'(A') \cap \mathcal{T}'(L) \neq \emptyset$. But $\mathcal{T}(A') \cap \mathcal{T}(L) = \emptyset$, contradicting the assumption that $\mathrm{int}(\mathcal{C},\mathcal{T}') \leq \mathrm{int}(\mathcal{C},\mathcal{T})$. Hence $A-L$ is empty, and there is indeed an edge e' in $\mathcal{T}$ that induces the X-split $L|(X-L)$.

Let s be an element not in X and set $X' = (X - L) \cup \{s\}$. We next define a collection $\mathcal{C}'$ of two-state characters on X', and two X'-trees $\mathcal{T}_s$ and $\mathcal{T}'_s$. For each character χ in $\mathcal{C}$, suppose the two states of χ are α and β, and define a character χ' on X' as follows. For all $x \in X - L$, if x is in the domain of χ, set $\chi'(x) = \chi(x)$; otherwise x is not in the domain of χ'. For s, set $\chi'(s)$ as follows:

(a) if $\chi^{-1}(\alpha) \cap L \neq \emptyset$ and $\chi^{-1}(\beta) \cap L = \emptyset$, then set $\chi'(s) = \alpha$;
(b) if $\chi^{-1}(\alpha) \cap L = \emptyset$ and $\chi^{-1}(\beta) \cap L \neq \emptyset$, then set $\chi'(s) = \beta$;
(c) if $\chi^{-1}(\alpha) \cap L \neq \emptyset$ and $\chi^{-1}(\beta) \cap L \neq \emptyset$, then either $\chi^{-1}(\alpha) \subseteq L$ in which case set $\chi'(s) = \beta$, or $\chi^{-1}(\beta) \subseteq L$ in which case set $\chi'(s) = \alpha$; and
(d) if $\chi^{-1}(\alpha) \cap L = \emptyset$ and $\chi^{-1}(\beta) \cap L = \emptyset$, then s is not in the domain of χ'.

The collection $\mathcal{C}'$ consists of the resulting collection of characters χ'. The X'-trees $\mathcal{T}_s$ and $\mathcal{T}'_s$ are obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by identifying the vertices that are labelled by elements of L, removing any loops, and then relabelling the resulting vertex s. We next show that $\mathcal{C}'$, $\mathcal{T}_s$, and $\mathcal{T}'_s$ satisfy the assumptions of their corresponding namesakes in the statement of the lemma.

Since $\mathcal{T}$ and $\mathcal{T}'$ display $\mathcal{C}$, it is easily seen that $\mathcal{T}_s$ and $\mathcal{T}'_s$ display $\mathcal{C}'$. Furthermore, since, by assumption, $\mathrm{int}(\mathcal{C}, \mathcal{T}') \leq \mathrm{int}(\mathcal{C}, \mathcal{T})$, it is also easily checked by the construction of $\mathcal{C}'$ that $\mathrm{int}(\mathcal{C}', \mathcal{T}'_s) \leq \mathrm{int}(\mathcal{C}', \mathcal{T}_s)$. Lastly, let f be an edge of $\mathcal{T}_s$, then the corresponding edge in $\mathcal{T}$ is strongly distinguished by some character $\chi \in \mathcal{C}$, and the associated character χ' in $\mathcal{C}'$ strongly distinguishes f. Hence, as $|X'| < |X|$, it follows by the inductive hypothesis that $\mathcal{T}'_s$ is a refinement of $\mathcal{T}_s$.

As $\mathcal{T}'_s$ is a refinement of $\mathcal{T}_s$ and $L|(X - L)$ is an X-split of $\mathcal{T}'$, we immediately deduce in case (i) that $\mathcal{T}'$ is a refinement of $\mathcal{T}$. Furthermore, in case (ii), for each leaf v adjacent to u, the edge $\{u, v\}$ is strongly distinguished in $\mathcal{T}$ and so there must be a character $\chi \in \mathcal{C}$ such that $\chi(a) \neq \chi(b)$ for all $b \in (L - \{a\})$, where $\{a\}$ is the label set of v. Hence $\{a\}$ is also the label set of a leaf in $\mathcal{T}'$, and as $\mathcal{T}'_s$ is a refinement of $\mathcal{T}_s$, we deduce that $\mathcal{T}'$ is a refinement of $\mathcal{T}$. This completes the proof of the lemma. □

Lemma 3.3. *Let $\mathcal{C}$ be a collection of two-state characters on X, and let $\mathcal{T}$ be an X-tree that displays $\mathcal{C}$ and is strongly distinguished by $\mathcal{C}$. Let $\mathcal{T}'$ be an X-tree that displays $\mathcal{C}$. If $\mathrm{int}(\mathcal{C}, \mathcal{T}) \leq \mathrm{int}(\mathcal{C}, \mathcal{T}')$, then $\mathcal{T}'$ is a refinement of $\mathcal{T}$.*

Proof. Suppose that $\mathcal{T}'$ is not a refinement of $\mathcal{T}$. Let $\mathcal{T}_{ph}$ and $\mathcal{T}'_{ph}$ be the two phylogenetic X-trees that are obtained from from $\mathcal{T}$ and $\mathcal{T}'$ respectively, by replacing each label of an internal vertex and each label of a multi-labelled leaf, by a leaf attached to the original vertex with a new edge and taking the same label. Now we observe the following.

(i) $\mathcal{T}_{ph}$ and $\mathcal{T}'_{ph}$ both display $\mathcal{C}$.
(ii) Since every edge of $\mathcal{T}$ is strongly distinguished by a character in $\mathcal{C}$, every internal edge of $\mathcal{T}_{ph}$ is also strongly distinguished by a character in $\mathcal{C}$.
(iii) Lastly, $\mathrm{int}(\mathcal{C}, \mathcal{T}_{ph}) \leq \mathrm{int}(\mathcal{C}, \mathcal{T}'_{ph})$. This can be seen by supposing that $\{(\chi_1, A_1), (\chi_2, A_2)\}$ is an edge in $\mathrm{int}(\mathcal{C}, \mathcal{T}_{ph})$, but not an edge in $\mathrm{int}(\mathcal{C}, \mathcal{T}'_{ph})$. Since $\{(\chi_1, A_1), (\chi_2, A_2)\}$ is an edge in $\mathrm{int}(\mathcal{C}, \mathcal{T}_{ph})$, it is also an edge of

$\text{int}(\mathcal{C}, \mathcal{T})$ and therefore, by our initial assumptions, also an edge in $\text{int}(\mathcal{C}, \mathcal{T}')$. Hence $\mathcal{T}'(A_1) \cap \mathcal{T}'(A_2) \neq \emptyset$. Now if $|A_1|, |A_2| \geq 2$, then the definition of $\mathcal{T}'_{ph}$ implies that $\mathcal{T}'_{ph}(A_1) \cap \mathcal{T}'_{ph}(A_2) \neq \emptyset$ and so $\{(\chi_1, A_1), (\chi_2, A_2)\}$ must be an edge in $\text{int}(\mathcal{C}, \mathcal{T}'_{ph})$; a contradiction. Hence, we may assume that either $|A_1| = 1$ or $|A_2| = 1$. Without loss of generality, assume that $|A_1| = 1$. But then A_1 is the label set of a leaf in $\mathcal{T}'_{ph}$ and so, as $A_1 \cap A_2$ is empty, $\mathcal{T}_{ph}(A_1) \cap \mathcal{T}_{ph}(A_2)$ is empty, contradicting the assumption that $\{(\chi_1, A_1), (\chi_2, A_2)\}$ is an edge in $\text{int}(\mathcal{C}, \mathcal{T}_{ph})$.

We may now apply [6, Corollary 3.5], which is the special case of Lemma 3.3 for when $\mathcal{T}$ and $\mathcal{T}'$ are both phylogenetic X-trees and every interior edge of $\mathcal{T}$ is strongly distinguished by $\mathcal{C}$, to $\mathcal{T}_{ph}$ and $\mathcal{T}'_{ph}$. Hence $\mathcal{T}'_{ph}$ is a refinement of $\mathcal{T}_{ph}$, and so, as $\mathcal{T}_{ph}$ is itself a refinement of $\mathcal{T}$, it follows that $\mathcal{T}'_{ph}$ is a refinement of $\mathcal{T}$. Since $\mathcal{T}'$ is not a refinement of $\mathcal{T}$, there is some X-split of $\mathcal{T}$ which is not an X-split of $\mathcal{T}'$. Furthermore, since $\mathcal{T}'_{ph}$ is a refinement of $\mathcal{T}$, it has an edge e inducing this X-split. By construction of $\mathcal{T}'_{ph}$, this X-split must be of the form $a|X - a$ for some element $a \in X$. Since $a|X - a$ is an X-split of $\mathcal{T}$ and $\mathcal{T}$ is strongly distinguished by $\mathcal{C}$, there is a leaf of $\mathcal{T}$ labelled a and there is a character $\chi \in \mathcal{C}$ which strongly distinguishes the adjacent edge. Again since $\mathcal{T}'_{ph}$ is a refinement of $\mathcal{T}$, it follows that χ strongly distinguishes e and hence that $\mathcal{T}'$ does not display χ. This contradicts the assumption that $\mathcal{T}'$ displays $\mathcal{C}$ and we conclude that $\mathcal{T}'$ is a refinement of $\mathcal{T}$. □

Proof of Theorem 3.1. Suppose that (i) and (ii) in the statement of Theorem 3.1 hold. By the definition of infers, it suffices to show that $\mathcal{C}$ identifies an X-tree if (i) is replaced by (i)′:

(i)′ there is an X-tree that displays $\mathcal{C}$ and is strongly distinguished by $\mathcal{C}$.

Let $\mathcal{T}$ be an X-tree as described by (i)′ and let $\mathcal{T}'$ be an X-tree such that $\text{int}(\mathcal{C}, \mathcal{T}')$ is the unique maximal element of $\mathcal{G}(\mathcal{C})$. Then, as $\mathcal{T}$ and $\mathcal{T}'$ satisfy the conditions of Lemma 3.3, $\mathcal{T}'$ is a refinement of $\mathcal{T}$. This implies that $\text{int}(\mathcal{C}, \mathcal{T}') \leq \text{int}(\mathcal{C}, \mathcal{T})$ and so, as $\text{int}(\mathcal{C}, \mathcal{T}')$ is the unique maximal element of $\mathcal{G}(\mathcal{C})$, $\text{int}(\mathcal{C}, \mathcal{T}) = \text{int}(\mathcal{C}, \mathcal{T}')$. Now, for any X-tree $\mathcal{T}''$ displaying $\mathcal{C}$, $\text{int}(\mathcal{C}, \mathcal{T}'') \leq \text{int}(\mathcal{C}, \mathcal{T})$. It now follows by Lemma 3.2 that $\mathcal{T}''$ is a refinement of $\mathcal{T}$, and so $\mathcal{C}$ identifies $\mathcal{T}$. □

4. Proof of Theorem 1.3

This section consists of the proof of Theorem 1.3. We will need the following lemmas, the first of which is established in [6], to enable us to apply Theorem 3.1 to this more general setting. Let $\mathcal{C}$ be a collection of characters on X. For a character $\chi : X' \to \{\alpha_1, \ldots, \alpha_n\}$ in $\mathcal{C}$, define characters

$$\chi_{i,j} : \chi^{-1}(\alpha_i) \cup \chi^{-1}(\alpha_j) \to \{\alpha_i, \alpha_j\}, \quad 1 \leq i < j \leq n,$$

by $\chi_{i,j}(x) = \chi(x)$, for $x \in \chi^{-1}(\alpha_i) \cup \chi^{-1}(\alpha_j)$. Consider the set of two-state characters $\mathcal{C}' = \bigcup_{\chi \in \mathcal{C}} \bigcup_{1 \leq i < j \leq n} \{\chi_{i,j}\}$ on X.

Lemma 4.1. *An X-tree $\mathcal{T}$ displays $\mathcal{C}$ if and only if $\mathcal{T}$ displays $\mathcal{C}'$.*

Lemma 4.2. *Let $\mathcal{T}$ be an X-tree that displays $\mathcal{C}$ (or, equivalently, displays $\mathcal{C}'$). If* $\text{int}(\mathcal{C}', \mathcal{T})$ *is a maximal element in $\mathcal{G}(\mathcal{C}')$, then* $\text{int}(\mathcal{C}, \mathcal{T})$ *is a maximal element in $\mathcal{G}(\mathcal{C})$.*

Proof. Suppose that $\text{int}(\mathcal{C}', \mathcal{T})$ is a maximal element in $\mathcal{G}(\mathcal{C}')$, but $\text{int}(\mathcal{C}, \mathcal{T})$ is not a maximal element in $\mathcal{G}(\mathcal{C})$. Then there is an X-tree $\mathcal{T}'$ that displays $\mathcal{C}$ such that $\text{int}(\mathcal{C}, \mathcal{T}) < \text{int}(\mathcal{C}, \mathcal{T}')$. By Lemma 4.1, $\mathcal{T}'$ displays $\mathcal{C}'$, and so $\text{int}(\mathcal{C}', \mathcal{T}')$ is an element of $\mathcal{G}(\mathcal{C}')$. We complete the proof by showing that $\text{int}(\mathcal{C}', \mathcal{T}) < \text{int}(\mathcal{C}', \mathcal{T}')$, thus contradicting the assumption that $\text{int}(\mathcal{C}', \mathcal{T})$ is a maximal element in $\mathcal{G}(\mathcal{C}')$.

For any two subsets $A \in \pi(\chi)$ and $B \in \pi(\chi')$ of X with $\chi, \chi' \in \mathcal{C}$, if $\mathcal{T}'(A) \cap \mathcal{T}'(B) = \emptyset$, then, as $\text{int}(\mathcal{C}, \mathcal{T}) < \text{int}(\mathcal{C}, \mathcal{T}')$, we have $\mathcal{T}(A) \cap \mathcal{T}(B) = \emptyset$. Moreover, there exist characters $\chi, \chi' \in \mathcal{C}$ and subsets $A \in \pi(\chi)$ and $B \in \pi(\chi')$ of X such that $\mathcal{T}'(A) \cap \mathcal{T}'(B) \neq \emptyset$ but $\mathcal{T}(A) \cap \mathcal{T}(B) = \emptyset$. It now follows by the definition of $\mathcal{C}'$ that $\text{int}(\mathcal{C}', \mathcal{T}) < \text{int}(\mathcal{C}', \mathcal{T}')$. □

Proof of Theorem 1.3. Suppose that $\mathcal{C}$ identifies an X-tree $\mathcal{T}$. Then, by definition, $\mathcal{T}$ displays $\mathcal{C}$ and, for every edge e of $\mathcal{T}$, there is a character on X inferred by $\mathcal{C}$ that strongly distinguishes e, namely the X-split of $\mathcal{T}$ induced by e. Thus (i) in the statement of Theorem 1.3 holds. Furthermore, since every X-tree displaying $\mathcal{C}$ is a refinement of $\mathcal{T}$, $\text{int}(\mathcal{C}, \mathcal{T}') \leq \text{int}(\mathcal{C}, \mathcal{T})$ for all $\mathcal{T}'$ displaying $\mathcal{C}$. Thus $\text{int}(\mathcal{C}, \mathcal{T})$ is the unique maximal element of $\mathcal{G}(\mathcal{C})$, and (ii) in the statement of Theorem 1.3 holds.

Now suppose that conditions (i) and (ii) in the statement of Theorem 1.3 hold. Let $\mathcal{T}$ be an X-tree that satisfies the properties in (i). Using Lemmas 4.1 and 4.2, we now show that conditions (i) and (ii) in the statement of Theorem 3.1 hold with "$\mathcal{C}$" replaced by "$\mathcal{C}'$". Using Lemma 4.1, it is easily seen that $\mathcal{T}$ displays $\mathcal{C}'$ and is strongly distinguished by a character inferred by $\mathcal{C}'$. Now suppose that there are two distinct maximal elements of $\mathcal{G}(\mathcal{C}')$. Call these elements G_1' and G_2', and let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two X-trees such that $\text{int}(\mathcal{T}_1, \mathcal{C}') = G_1'$ and $\text{int}(\mathcal{T}_2, \mathcal{C}') = G_2'$. By Lemma 4.2, $\text{int}(\mathcal{T}_1, \mathcal{C})$ and $\text{int}(\mathcal{T}_2, \mathcal{C})$ are both maximal elements of $\mathcal{G}(\mathcal{C})$. We now show that these last two graphs are distinct, thus contradicting our original assumption that there is a unique maximal element of $\mathcal{G}(\mathcal{C})$.

Since G_1' and G_2' are distinct, there is an edge $\{(\chi_A', A), (\chi_B', B)\}$ of G_1' that is not an edge of G_2'; that is, $\mathcal{T}_1(A) \cap \mathcal{T}_1(B) \neq \emptyset$, but $\mathcal{T}_2(A) \cap \mathcal{T}_2(B) = \emptyset$. Since, by construction of $\mathcal{C}'$ there are characters χ_A and χ_B in $\mathcal{C}$ such that $A \in \pi(\chi_A)$ and $B \in \pi(\chi_B)$, it follows that $\{(\chi_A, A), (\chi_B, B)\}$ must be an edge of $\text{int}(\mathcal{T}_1, \mathcal{C})$ and not of $\text{int}(\mathcal{T}_2, \mathcal{C})$. This implies that $\text{int}(\mathcal{T}_1, \mathcal{C})$ and $\text{int}(\mathcal{T}_2, \mathcal{C})$ are distinct thereby contradicting that there is a unique maximal element of $\mathcal{G}(\mathcal{C})$. Hence there is a unique maximal element of $\mathcal{G}(\mathcal{C}')$.

We now deduce by Theorem 3.1 that $\mathcal{C}'$ identifies $\mathcal{T}$. This in turn implies by Lemma 4.1 that $\mathcal{C}$ identifies $\mathcal{T}$, thus completing the proof of Theorem 1.3. □

5. Proof of Theorem 1.5

This section consists of the proof of Theorem 1.5. To this end, we first prove two lemmas.

Lemma 5.1. *Let $\mathcal{C}$ be a collection of characters on X. Let G be a chordal completion of* $\text{int}(\mathcal{C})$ *and let $\mathcal{C}'$ be the subset of $\mathcal{C}$ consisting of the broken characters in G. Then there is an X-tree that displays $\mathcal{C} - \mathcal{C}'$.*

Proof. Using the maximal clique tree construction (see [4] or [7] for details), there exists a tree T' whose vertex set $\mathcal{K}$ is the set of maximal cliques of G and, for each vertex (χ, A) in G, the subgraph of T' induced by the elements of $\mathcal{K}$ containing (χ, A) is a subtree of T'. Observe that, if x is an element of X, then since $\text{int}(\mathcal{C})$ is a subgraph of G, the vertices $V_x = \{(\chi, A) \in V(G) : x \in A, \chi \in \mathcal{C}\}$ in G form a clique.

Define $\phi : X \to \mathcal{K}$ to be a map with the property that, for each element x in X, $\phi(x)$ is a maximal clique in G that contains V_x. Now define T to be the tree obtained from T' by suppressing all vertices of degree two that are not contained in the image of ϕ. It is easily checked that all degree-one vertices of T' are contained in the image of ϕ, and so $\mathcal{T} = (T; \phi)$ must be an X-tree.

To complete the proof, we show that if χ is an element of $\mathcal{C}$ and is not broken in G, then $\mathcal{T}$ displays χ. Suppose $A_1, A_2 \in \pi(\chi)$ and let T_1' and T_2' be the subtrees of T' induced by the elements of $\mathcal{K}$ containing (χ, A_1) and (χ, A_2), respectively. Since χ is not broken in G, no maximal clique in G can contain both (χ, A_1) and (χ, A_2). By the construction of T', this implies that T_1' and T_2' do not share a vertex. Since T is obtained from T' by suppressing degree-two vertices, it follows that $\mathcal{T}(A_1)$ and $\mathcal{T}(A_2)$ cannot share a vertex either. Hence if $\chi \in \mathcal{C}$ is not broken in G, then $\mathcal{T}$ displays χ. □

Lemma 5.2. *Let $\mathcal{C}$ be a collection of characters on X and let $\mathcal{T}$ be an X-tree. Let $\mathcal{C}'$ be the subset of characters in $\mathcal{C}$ that are displayed by $\mathcal{T}$. Then* $\text{int}(\mathcal{C}, \mathcal{T})$ *is a chordal completion of* $\text{int}(\mathcal{C})$ *in which the broken characters are precisely the elements in $\mathcal{C} - \mathcal{C}'$.*

Proof. By [6, Corollary 2.2], $\text{int}(\mathcal{C}, \mathcal{T})$ is chordal. Therefore, as the edge set of $\text{int}(\mathcal{C})$ is a subset of the edge set of $\text{int}(\mathcal{C}, \mathcal{T})$, it follows that $\text{int}(\mathcal{C}, \mathcal{T})$ is a chordal completion of $\text{int}(\mathcal{C})$. Let χ be a character in $\mathcal{C}$. If $\mathcal{T}$ displays χ, then, for all distinct $A_1, A_2 \in \pi(\chi)$, the vertices (χ, A_1) and (χ, A_2) in $\text{int}(\mathcal{C}, \mathcal{T})$ are not joined by an edge in $\text{int}(\mathcal{C}, \mathcal{T})$. Hence χ is not broken in $\text{int}(\mathcal{C}, \mathcal{T})$.

Now suppose that $\mathcal{T} = (T; \phi)$ does not display χ. Then there exist B_1 and B_2 in $\pi(\chi)$ such that $\mathcal{T}(B_1) \cap \mathcal{T}(B_2)$ is non-empty. By the definition of $\text{int}(\mathcal{C}, \mathcal{T})$, this implies that (χ, B_1) and (χ, B_2) are joined by an edge in $\text{int}(\mathcal{C}, \mathcal{T})$, and so χ is broken in $\text{int}(\mathcal{C}, \mathcal{T})$. This completes the proof of the lemma. □

Proof of Theorem 1.5. Suppose that $\mathcal{C}'$ is a maximum-sized compatible subset of $\mathcal{C}$. Then there is an X-tree $\mathcal{T}$ that displays $\mathcal{C}'$. By Lemma 5.2, $\text{int}(\mathcal{C}, \mathcal{T})$ is a

chordal completion of int($\mathcal{C}$) and in which the broken characters in $\mathcal{C}$ are exactly the elements of $\mathcal{C} - \mathcal{C}'$. To see that $\mathcal{C} - \mathcal{C}'$ is a minimum-sized broken subset of $\mathcal{C}$, suppose there exists some broken subset $\mathcal{C}''$ of $\mathcal{C}$ with $|\mathcal{C}''| < |\mathcal{C} - \mathcal{C}'|$. Then, by Lemma 5.1, there is an X-tree that displays $\mathcal{C}-\mathcal{C}''$. But $|\mathcal{C}-\mathcal{C}''| > |\mathcal{C}'|$, contradicting the maximality of $\mathcal{C}'$. Thus $\mathcal{C} - \mathcal{C}'$ is a minimum-sized broken subset of $\mathcal{C}$.

Now suppose that $\mathcal{C} - \mathcal{C}'$ is a minimum-sized broken subset of $\mathcal{C}$. Then, by Lemma 5.1, there exists an X-tree that displays $\mathcal{C}'$, and so $\mathcal{C}'$ is compatible. To see that $\mathcal{C}'$ is a maximum-sized compatible subset of C, suppose there exists a compatible subset $\mathcal{C}''$ of $\mathcal{C}$ with $|\mathcal{C}''| > |\mathcal{C}'|$. Then there is an X-tree T' that displays $\mathcal{C}''$. By Lemma 5.2, int$(\mathcal{C}, T')$ is a chordal completion of int($\mathcal{C}$) in which the broken characters are exactly the elements of $\mathcal{C} - \mathcal{C}''$. This implies that $\mathcal{C}$ has a broken subset of size $|\mathcal{C} - \mathcal{C}''| < |\mathcal{C} - \mathcal{C}'|$; a contradiction. Hence $\mathcal{C}'$ is a maximum-sized compatible subset of $\mathcal{C}$. □

REFERENCES

[1] H. L. Bodleander, M. R. Fellows, T. J. Warnow, Two strikes against perfect phylogeny, in: Proceedings of the International Colloquium on Automata, Languages and Programming. Lecture Notes in Computer Science, Vol. 623, Springer, Berlin, 1993, 273-283.

[2] P. Buneman, A characterization of rigid circuit graphs, *Discrete Math.* **9** 205-212 (1974).

[3] P. Daniel, Supertree methods: Some New Approaches, MSc Thesis, University of Canterbury, 2004.

[4] M. C. Golumbic, Algorithmic graph theory and perfect graphs, Academic Press, New York, 1980.

[5] C. A. Meacham, Theoretical and computational considerations of the compatibility of qualitative taxonomic characters, in: J. Felsenstein (Ed.), Numerical Taxonomy, NATO ASI Series Vol. G1, Springer, Berlin, 1983, 304-314.

[6] C. Semple and M. Steel, A characterization for a set of partial partitions to define an X-tree, *Discrete Math.* **247** 169-186 (2002).

[7] C. Semple and M. Steel, Phylogenetics, Oxford University Press, 2003.

[8] M. Steel, The complexity of reconstructing trees from qualitative characters and subtrees, *J. Classification* **9**(1) 91-116 (1992).

[9] T. Warnow, Constructing phylogenetic trees efficiently using compatibility criteria, *New Zealand Journal of Botany* **31** 329-248 (1993).

DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

E-mail address: m.bordewich@math.canterbury.ac.nz, c.semple@math.canterbury.ac.nz

DEPARTMENT OF BIOMETRY AND ENGINEERING, THE SWEDISH UNIVERSITY OF AGRICULTURAL SCIENCES, BOX 7013, 750 07 UPPSALA, SWEDEN and THE LINNAEUS CENTRE FOR BIOINFORMATICS, UPPSALA UNIVERSITY, BMC, BOX 598, 751 24 UPPSALA, SWEDEN

E-mail address: katharina.huber@lcb.uu.se